TRABAJO PRÁCTICO Nº1

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION (NCBI) Y EL ANALISIS DE SECUENCIAS

Con el enorme incremento en número de secuencias nucleotídicas y de proteínas que se fue dando desde 1980, fue necesaria la creación de bases de datos para guardar dicha información, de tal forma que todos los investigadores y público en general tengan acceso libre a dichas secuencias. A su vez, resulta importante que los investigadores puedan incluir en las bases de datos las nuevas secuencias generadas. A principios de los 80's, Europa, Japón y EEUU crearon independientemente los tres mayores centros de bioinformática del mundo. Estos centros forman hoy la International Sequence Database (INSD) y cooperan en el resguardo de las secuencias y también en la preparación de herramientas para el análisis de las mismas. Los tres centros son los siguientes:

1. The National Center for Biotechnology Information (NCBI) /GenBank: http://www.ncbi.nlm.nih.gov/Genbank

2. The European Bioinformatics Institute (EBI)/The European Molecular Biology Laboratory (EMBL)/ EMBL-Bank, SWISS-PROT, TrEMBL, interPro, E-MSD, ENSEMBLE: http://www.ebi.ac.uk

3. National Institute of Genetics (NIG)-Center for Information Biology (CIB)/ DNA Data Bank of Japan (DDBJ): http://www.ddbj.nig.ac.jp

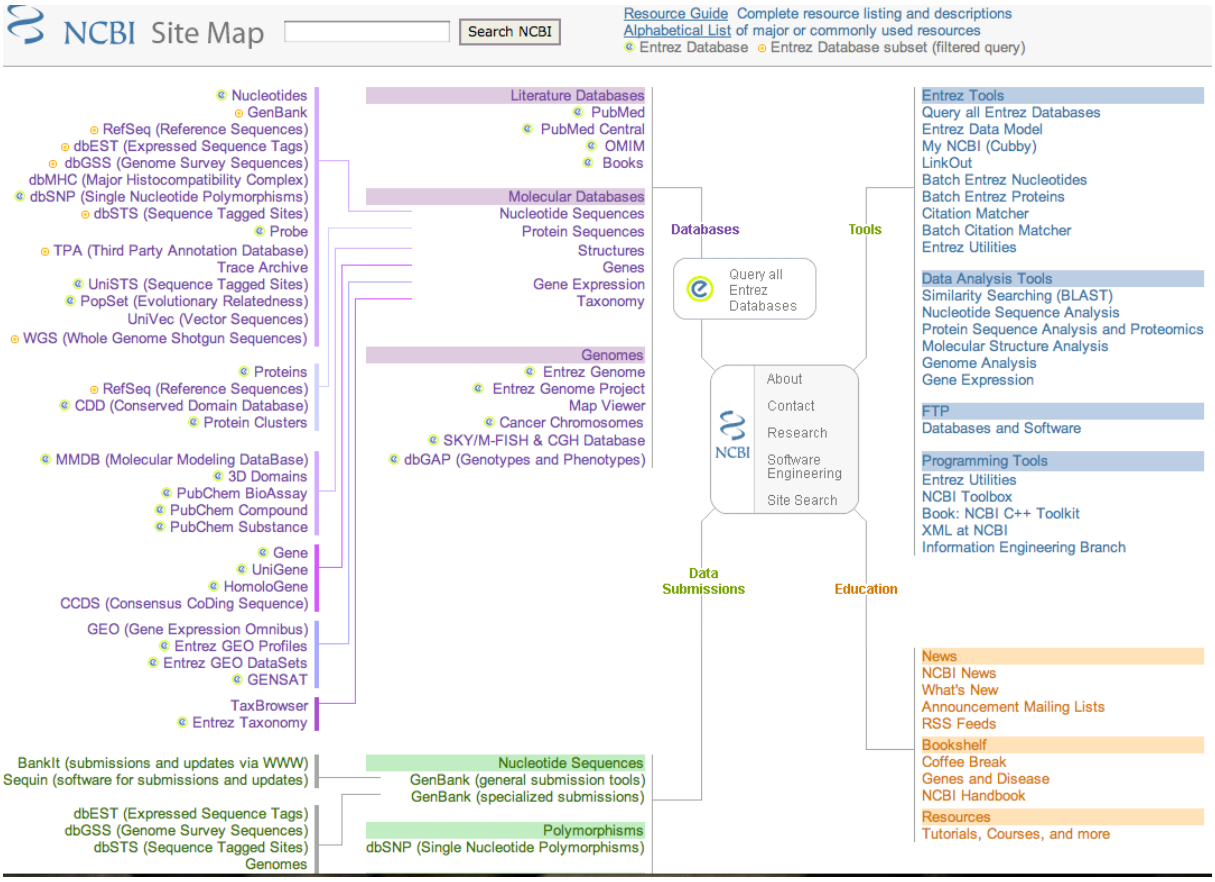
Las nuevas secuencias depositadas en una de estas tres bases de datos son copiadas diariamente a las otras dos, y así las tres bases de datos son actualizadas cada día y mantienen la misma información. De esta forma, ante cualquier desastre que pudiera ocurrir a alguno de estos centros, tan valiosa información estaría a salvo en los otros dos y no se perdería. La información que se encuentra en cada base de datos es prácticamente la misma, aunque la forma de presentación y las herramientas para el análisis de las mismas varían en cada centro. Aquí nos focalizaremos en el uso de las bases de datos ofrecidas por NCBI.

NCBI fue creado en 1988 y desde entonces está encargado de proveer principalmente dos servicios: 1) guardar la literatura e información de secuencias producida en todo el mundo en bases de datos que incluyen base de datos bibliográficas, moleculares y de genomas; 2) desarrollar y ofrecer herramientas que faciliten el acceso y análisis de la información guardada en dichas bases de datos.

Actividad 1. Recorrido general de NCBI y Bases de datos Bibliográficas.

Visite el sitio: http://www.ncbi.nlm.nih.gov/

El siguiente es un mapa del sitio NCBI para orientar al usuario sobre las bases de datos existentes, las relaciones entre las mismas y los servicios y recursos ofrecidos por NCBI.



1. Realice una búsqueda de libros relacionados con algas ("algae")/protistas ("protistology") que estén disponibles para el público en la base de datos de literatura (**Books**). Cuántos encontró?

2. Realice la búsqueda de artículos en PMC (PubMed Central) utilizando el término: "secondary endosymbiosis".
Note que las búsquedas son **en ingles** y las frases van **entre comillas**.

a. Cuántos resultados obtuvo?

b. Descargue el primer artículo. Cuál es el título, en qué revista se publicó y en qué año?

3. Realice la misma búsqueda pero esta vez utilice PubMed.

a. Cuántos resultados obtuvo?

b. Descargue el primer artículo.

Actividad 2. Genomas secuenciados

Visite el sitio: http://www.ncbi.nlm.nih.gov/
Clickee en *Resource List* a la izquierda. Allí aparecerán términos en orden alfabético relacionados con los recursos ofrecidos por NCBI. Si clickea en cada uno de los términos encontrará una corta descripción de dicho recurso.

1. Dentro del índice alfabético, ingrese a "**Genome**". Luego ingrese a: **Organelles**. Explore:

a) ¿Cuántos genomas de mitocondrias de algas verdes (Plants -> Viridiplantae -> Chlorophyta) han sido completamente secuenciados? Nombre alguno e indique el tamaño.

b) Y cuántos cloroplastos de algas verdes? Nombre alguno e indique el tamaño.

Actividad 3. Uso de la base de datos de Taxonomía de NCBI

1. Abrir en el Block de Notas el archivo **gen.fas** y responda:
   a. Averiguar la taxonomía de cada una de las especies de plantas bajo estudio. Utilice la base de datos de taxonomía de NCBI: http://www.ncbi.nlm.nih.gov/taxonomy. Anotar para cada taxón, todos los grupos taxonómicos a los que pertenece a partir de Viridiplantae.

   b. ¿Cuál de las especies bajo estudio se podría considerar un outgroup según sus conocimientos previos?

Actividad 4. BLAST

En este ejercicio exploraremos una de las herramientas más usadas y más útiles de NCBI para el análisis de secuencias: Basic Local Alignment Search Tool (BLAST). BLAST es una familia de algoritmos diseñados con propósitos diversos y aquí exploraremos alguno de ellos. Esta herramienta permite al usuario, por ejemplo, comparar dos secuencias entre si o analizar la similitud de una secuencia elegida contra todas las secuencias mantenidas en una base de datos de NCBI, ya sea de nucleótidos o de proteínas.

BLASTn se utiliza para comparar una secuencia de nucleótidos contra las bases de datos de nucleótidos en NCBI (i.e. nucleótidos vs nucleótidos).

BLASTx toma una secuencia de nucleótidos a comparar, la traduce a una secuencia proteica y la compara contra las bases de datos de proteínas (i.e. nucleótidos traducidos vs proteínas).

1. Dado que el gen bajo estudio codifica para proteínas, utilice **BLASTx** para responder los siguientes interrogantes. Para ello incluya como "query" (secuencia para iniciar búsqueda) cualquiera de las secuencias dentro del archivo gen.fas. Para acortar el tiempo de búsqueda, restrinja la búsqueda a "angiosperms taxid:3398"

    a. Cuál es el gen que está analizando? Qué proteina codifica?

    b. en qué compartimiento celular se encuentra codificado (núcleo / cloroplasto / mitocondria)? Cuál es la función?

2. Obtener *secuencias homólogas* al gen bajo estudio de algas verdes para análisis comparativos y filogenéticos que realizaremos en futuros trabajos prácticos.
    a. Realice una búsqueda usando **BLASTn** a partir de una de las secuencias del archivo gen.fas, restringiendo la búsqueda a las algas verdes (Charophyceae=taxid 304574 y Chlorophyceae=taxid 3166).

    b. Obtener las secuencias *de nucleótidos* del gen bajo estudio de **al menos 2 algas verdes** (una carófita y una clorófita). Nota: si en los resultados (hits) del BLAST encuentra genomas completos, ingrese al genoma y busque el nombre del gen bajo estudio. Descargue el CDS (coding DNA sequence), que no incluirá intrones que pudieran estar presentes.

    c. Copiar ambas secuencias, indicando a que taxón pertenecen, en un archivo de texto del Block de notas. El archivo debe tener formato **FASTA** (el mismo formato que el archivo gen.fas).

    d. Si no encontró ningún resultado, realizar la búsqueda utilizando **BLASTx**. Por qué piensa que pueden no encontrarse secuencias similares con **BLASTn**?

    e. Registrar todos los grupos taxonómicos a los que pertenece cada uno, a partir de Viridiplantae.

3. Realice una búsqueda con **BLASTx** del gen bajo estudio en Bacterias.
    a. Cuál es la bacteria con la secuencia más similar a la de las plantas?

    b. A qué grupo de Bacteria pertenece?

    c. Interprete los resultados obtenidos.